МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
КАФЕДРА ОБЩЕСТВЕННОГО ЗДОРОВЬЯ И ЗДРАВООХРАНЕНИЯ

БИОМЕДИЦИНСКАЯ СТАТИСТИКА

BIOMEDICAL STATISTICS

Практикум

Минск БГМУ 2025

УДК 61:311.4(076.5)-054.6
ББК 51.1(2)-923
Б63

Рекомендовано Научно-методическим советом университета в качестве практикума 16.04.2025 г., протокол № 8

А в т о р ы: И. Н. Мороз, С. В. Власова, С. В. Куницкая, А. Н. Черевко, Л. А. Наумова, М. А. Лях, А. Н. Стецик, Е. С. Игумнова

Р е ц е н з е н т ы: д-р мед. наук, проф., зав. каф. общественного здоровья и здравоохранения с курсом ФПКиПК Витебского государственного ордена Дружбы народов медицинского университета В. С. Глушанко; каф. общественного здоровья и здравоохранения с курсом ФПКиП Гомельского государственного медицинского университета

Биомедицинская статистика = Biomedical statistics : практикум / И. Н. Мороз, Б63 С. В. Власова, С. В. Куницкая [и др.]. – Минск : БГМУ, 2025. – 52 с.

ISBN 978-985-21-1951-1.

Рассматриваются основные понятия и методы, используемые в медицинской статистике. Содержит разноуровневые задания в виде таблиц, текстов, схем. Информация предлагается как краткий конспект, что позволит студенту систематизировать полученные на лекциях знания и отработать практические навыки по биомедицинской статистике. Задания предназначены как для индивидуальной работы при подготовке к темам, так и для работы в группе на занятиях.

Предназначен для студентов 2-го курса медицинского факультета иностранных учащихся, обучающихся на английском языке по специальности «Лечебное дело».

УДК 61:311.4(076.5)-054.6
ББК 51.1(2)-923

ISBN 978-985-21-1951-1

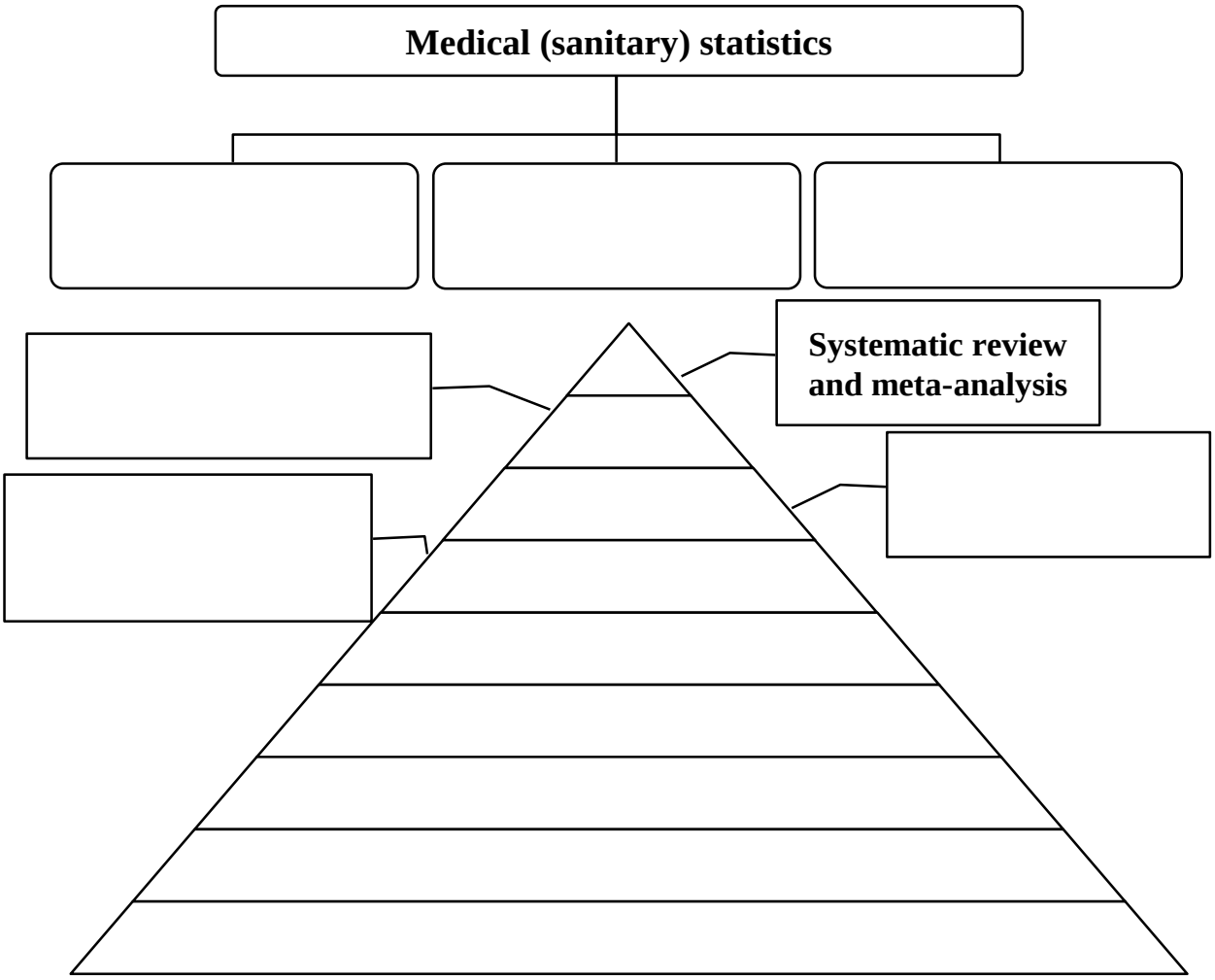
LEGEND

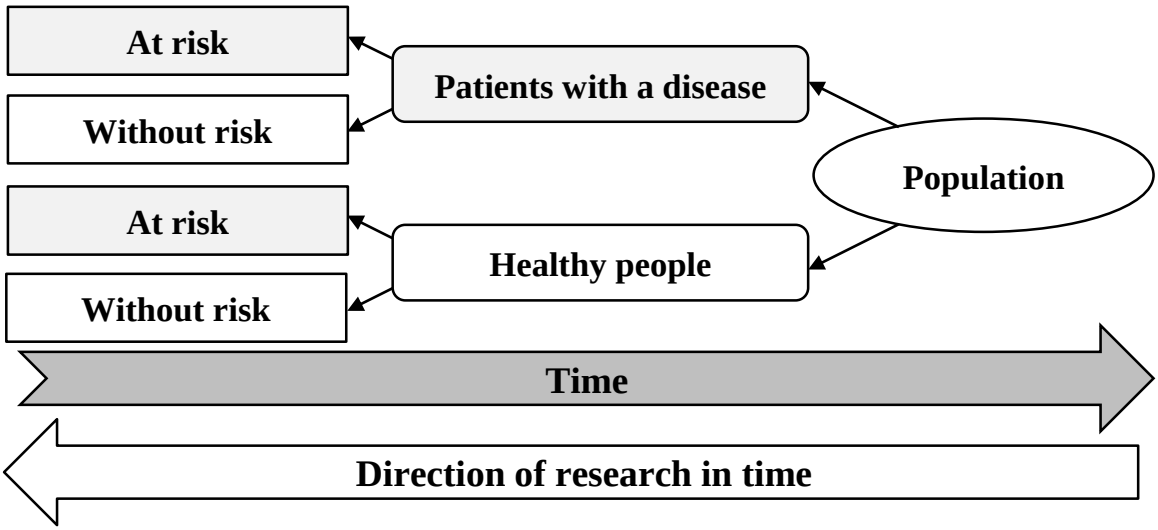
M — arithmetic mean
IQR — interquartile range
Me — median
Mo — mode
Q — quartile
A — range
σ — sample standard deviation
SD or std dev — standard deviation
Cv — coefficient of variation
σ^2 — dispersion
d — deviation
V — variable in the variation series
r_{xy} — Pearson correlation coefficient
ρ — coefficient of rank correlation
R — regression coefficient
P — Relative value
U — Mann–Whitney U Test
χ^2 — Chi-Square-test
RR — relative risk
AR — attributable risk
RRR — relative risk reduction

ORGANIZATION OF STATISTIC RESEARCH.
GRAPHIC IMAGES IN STATISTICS

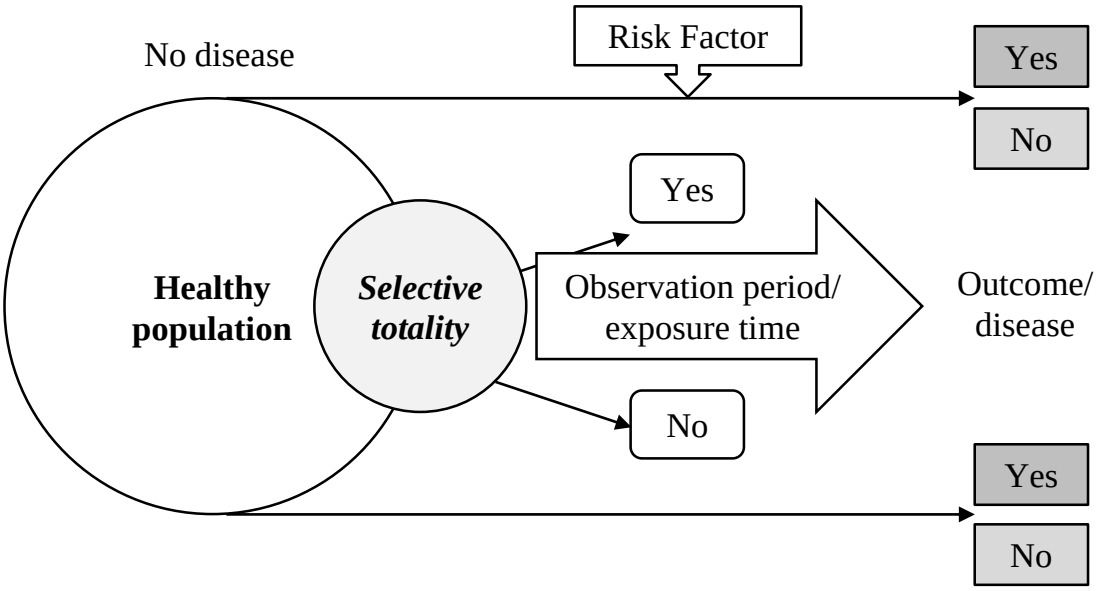
The term "statistics" (from Latin) word "status" means "_____".

Fill in the evidence pyramid with a study name.





This is an example _____ of research.



This is an example _____ of research.

STATISTIC RESEARCH STAGES

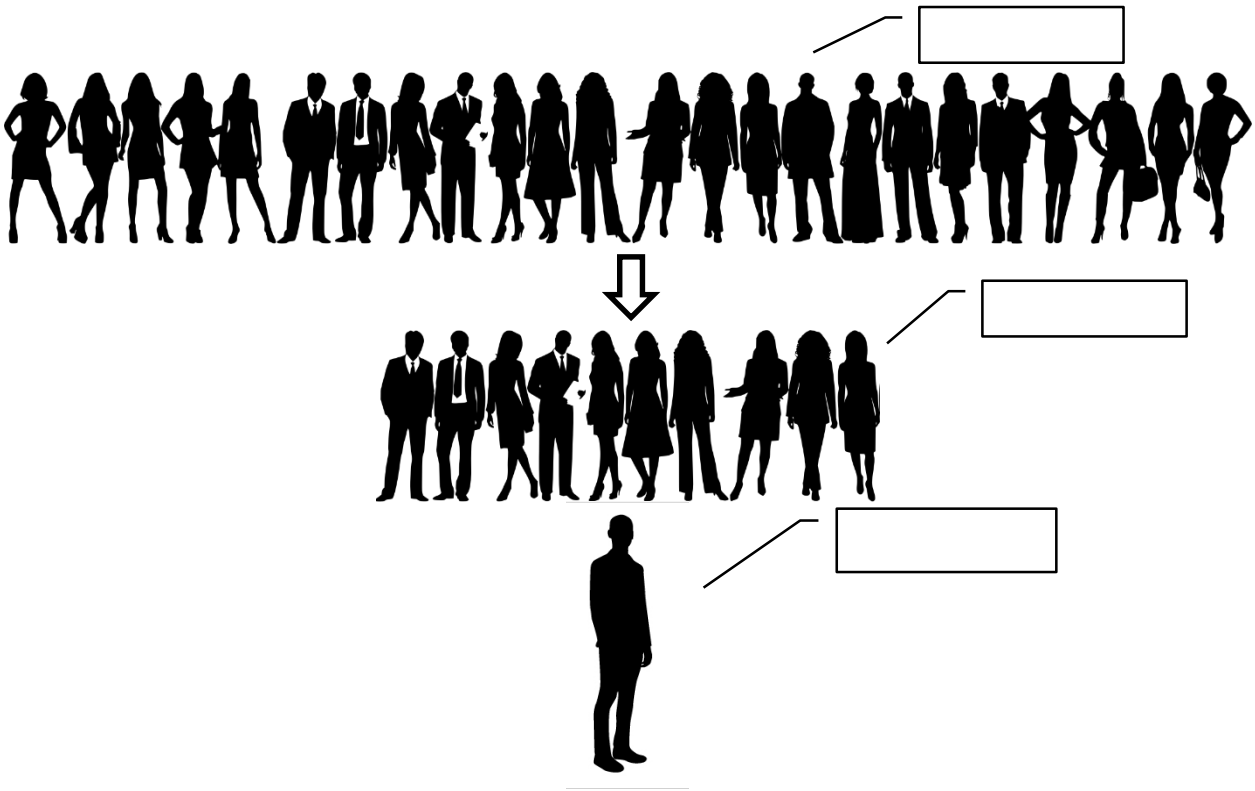
Fill the gaps.

1. _____
2. _____
3. _____
4. _____
5. _____

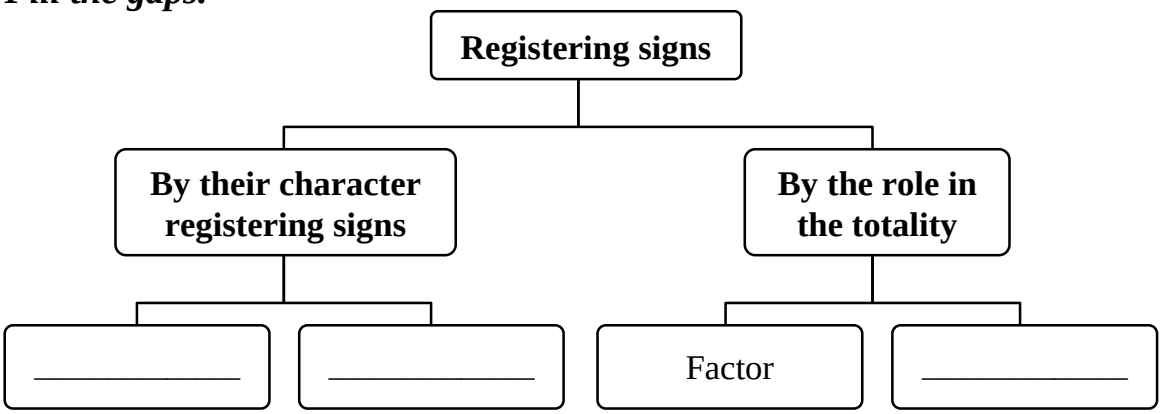
Observation unit (computation unit) is _____
_____
_____

Statistic population _____
_____

From a statistical point of view we can study **(fill the gaps)**.



Fill the gaps.



Fill the gaps.

Unit of _____

Registering signs	Character of registering signs
Age	*quantitative signs*
Sex	
Nosology form	
Severity of the patient's condition	
Level of hemoglobin	
Length of temporary disability	

Look at the table maquette, determine the type of tables presented:

1. Table 1 is _____

2. Table 2 is _____

3. Table 3 is _____

Table 1

Distribution of smoking students by faculties (in abs. numbers)

Department	Total number of students
General Medicine	
Preventive Medicine	
General Medicine and Diagnostics	
Total	

Table 2

Distribution of students of different faculties by gender, age of onset of smoking (in abs. numbers)

Department	sex		Age of onset of smoking (year)			Total
	m	f	up to 14	15–18	Adult	
General Medicine						
Preventive Medicine						
General Medicine and Diagnostics						
Total						

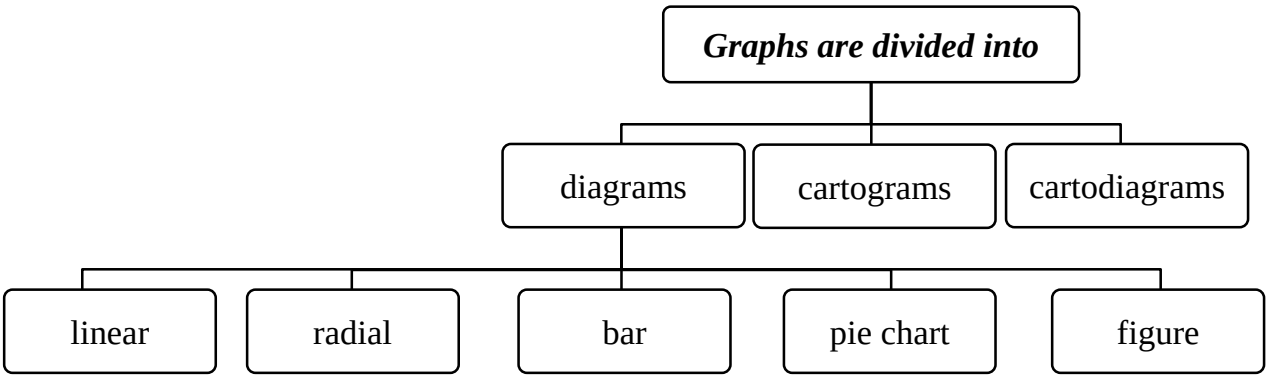
Table 3

Distribution of students of different faculties by sex and number of cigarettes smoked per day (in abs. numbers)

Department	Number of cigarettes smoked by students per day									Total		
	no more than half a pack			more than half a pack or a pack			more than 1 pack					
	m	f	both	m	f	both	m	f	both	m	f	both
General Medicine												
Preventive Medicine												
General Medicine and Diagnostics												
Total												

Grouping is the distribution of the collected material according to attributive or quantitative sign (typological or variable).

Graphic images. Graphs are used to visualize the results of scientific research.



_____ ***chart*** are used to show phenomenon dynamics. With the help of linear diagrams it is expedient to show indices dynamics of population movement, morbidity, network of medical establishments, etc.

_____ ***chart*** is based on the system of polar coordinates in showing phenomenon dynamics within a closed period of time (24 hours, a week, a year). Seasonal variations of infectious morbidity, daily variations of calls to the First Aid Station, fluctuations of discharged patients number and the admitted ones to hospitals on particular days of the week and so on.

_____ ***diagram*** allows to represent the dynamics of a phenomenon, as well as the structure of a phenomenon in a certain territory.

_____ ***chart*** is used to show the phenomenon structure (structure of morbidity or structure of population mortality causes, etc.).

Indicate which illustration shows the cartogram and what is the basis for its construction.

Fig. 1. World population

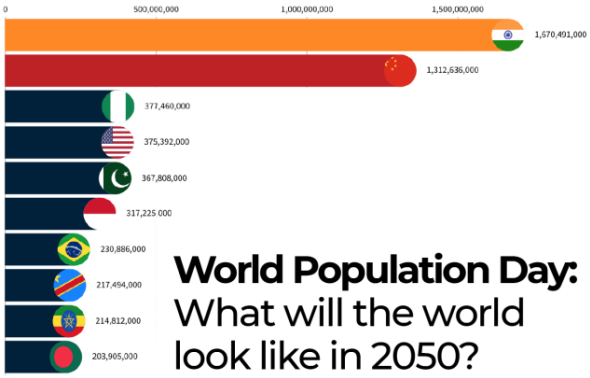


Fig. 2. Population forecast for different countries by 2050

_____

_____

_____

_____

_____

TASKS

Make a study plan, determine the main accounting feature and additional registration characteristics. Write separately qualitative signs (indicate whether there are ordinal signs) and quantitative characteristics. On the basis of the proposed task, make up the tables: a simple table, a group table, a combinational table.

Task 1

There was a study of hospital lethality for the year. The following signs were studied: age of patients (18–22, 23–27, 28–32, 33–37, 38–42, 43–47, 48–52, 53–57, 58–62, 63 years and older), sex (male, female), hospitalization from the onset of the disease (up to 12 hours, more than 12 hours), the class of the disease (vascular diseases, respiratory diseases, gastrointestinal diseases, trauma, other diseases), the severity of the disease (patient's mild condition, patient's condition of gravity, severe condition of the patient).

Task 2

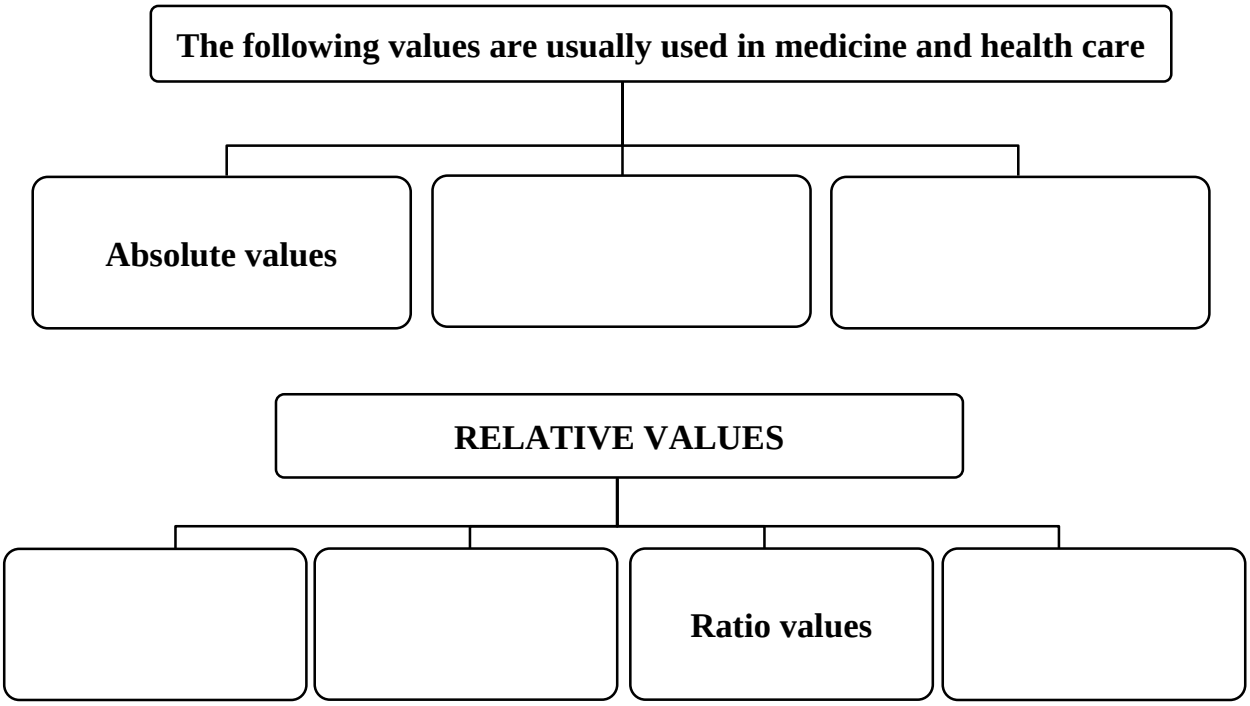
The study of childhood traumatism lasted for 1 year. The following signs were examined: the age of the child (age 1–2 years, age 3–4 years, age 5–6 years), the sex of the child (male, female), residence (urban, rural), attending the kindergarten (yes, no), localization of trauma (head, upper limbs, lower limbs, abdomen and thorax, spine, combined trauma).

Task 3

The study of the immune status of the children from the zone of ecological disaster was carried out for 5 years. The following signs have been analyzed: sex (male, female), age: early childhood period (from 1 year to 3 years); first period of childhood (from 3 to 7 years) — preschool age; second period of childhood (primary school age) — girls from 7 to 11 years, boys from 7 to 12 years; senior school period — girls from 12 years, boys — from 13 to 18 years, place of residence (urban area, countryside), presence genetic disorders (yes, no), presence of allergic diseases (yes, no), immune status (normal, violation of humoral immunity disorder deficiency, violation of cellular immunity, combination of violations of humoral and cellular immunity).

STATISTIC VALUES

Fill the gaps.



ABSOLUTE VALUES

Fill the gaps.

In some cases these values may be used to analyze the phenomenon:

– in the study of _____;

– if the exact _____ of the phenomenon is required;

– if it is necessary _____ of the studied phenomenon.

However, the absolute values can not be used to compare the data of several studies. The relative values and the mean values are required for this purpose.

RELATIVE VALUES

Fill the gaps.

Intensive value characterizes the frequency, level and the intensity of the phenomena in the environment, producing this phenomenon.

The equation for calculating the intensive values is (*fill the gap*).

$$\textbf{Intensive value} = \frac{\text{\_\_\_\_\_}}{\text{size of population, connected with phenomenon}} * \text{base.}$$

For example: _____

_____

Extensive value is the index of the phenomenon structure. It characterizes the distribution of the phenomenon into its component parts.

The equation for calculating the extensive values is (*fill the gap*).

$$\textbf{Extensive value} = \frac{\text{phenomenon part}}{\text{\_\_\_\_\_}} * 100\ (\%).$$

For example: _____
_____

Ratio value characterizes the ratio of two statistic sets, which are not related to each other and only logically correlated.

The equation for calculating ratio values is (*fill the gap*).

$$\textbf{Ratio value} = \frac{\text{\_\_\_\_\_}}{\text{\_\_\_\_\_}} * 10\ 000\ (1000).$$

For example: _____
_____

Obvious value is the index, which is used to characterize phenomenon changing in dynamics. This index is used also in cases, where it is necessary to show the direction of the process, trends without showing the level or the actual size of the phenomenon.

The equation for calculating the obvious values is (*fill the gap*).

$$\textbf{Obvious value} = \frac{\text{phenomenon}}{\text{another phenomenon is taken as \_\_\_\_\_}} * \text{\_\_\_\_\_}\ (\text{\_\_\_\_\_}).$$

For example: _____
_____

The result of calculation the relative value can be expressed as **(fill the gaps)**:

– per cent (%), if the basis is taken as _____.

– promille (‰), if the base is taken as _____.

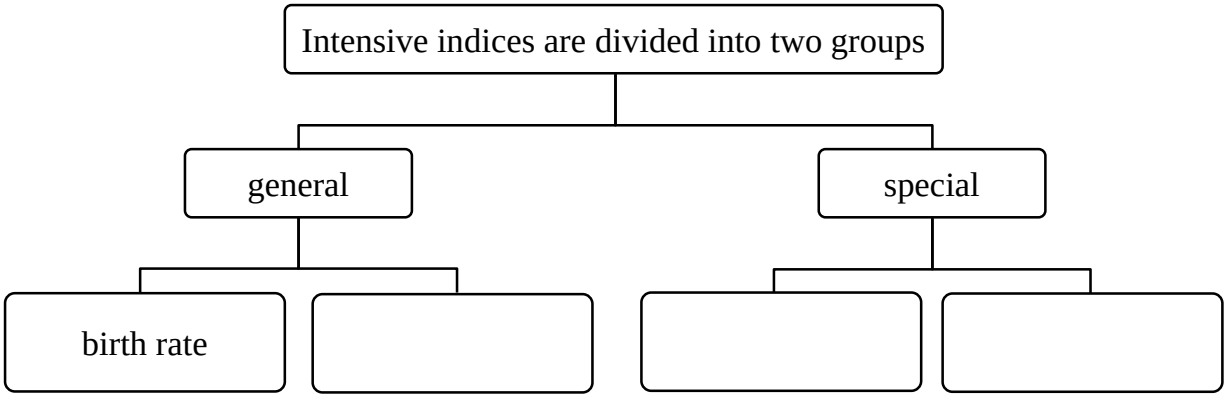
– prodecimille ($\%_{oo}$), if the base is taken as _____.

– prosantimille ($\%_{ooo}$), if the base is taken as _____.

Match the corresponding base and indicators.

The base	Examples of relative value
1000	is used to calculate the incidence
100	is used to calculate disability index
10 000	is used to calculate demographic indicators, such as births, deaths
100 000	is used to calculate index of hospital mortality

Fill the gaps with examples.



TASKS

Task 1

1. Calculate the relative values.
2. Specify the type of each relative value.
3. Graphically represent birth and mortality rates.
Material for calculating.

Table 1

Population, the number of births and deaths in A

Year	Average population	Live births	Deaths
Last year	70 496 000	1 253 912	408 566
Report year	75 149 000	1 382 229	422 133

Table 2

Distribution of the population in A

Population	Number of people
Children aged 0 to14	17 561 778
People of fertile age 15 to 49	45 174 366
People over 50 years and more	12 412 856

Task 2

1. Calculate the relative values.
2. Specify the types of each relative value.
3. Graphically represent structure of the population in B.
Material for calculating.

Table 1

Population, the number of births and deaths in B

Year	Average population	Live births	Deaths
Last year	9 485 500	116 975	110 510
Repor yeart	9 501 000	117 812	119 812

Table 2

Distribution of the population by sex for the reporting year in City B

Population	Number of people
Male	4 379 961
Female	5 121 039
Total	9 501 000

AVERAGES

The use of average values in medicine and health care

The average values are widely used in the daily work of medical workers. They are used to characterize the physical development:

_____

_____

The average values are used to evaluate the patient's condition by analyzing physiological, biochemical changes in the body:

_____

_____

The average values used while analyzing the activity of health care organizations:

_____

_____

The average values used to analyze the work of doctor:

_____

_____

Variational series

Variational series — an arrangement of homogeneous statistical values that characterize the same _____ accounting trait which differ from each in size and are arranged in a particular order (_____ _____).

These are the elements of variation series:

– _____ (V) — the numeric value of the studied changing the ***quantitative signs***;

– _____ (p) — indicate ***how often*** there is a version of the series;

– *the total number of observations* (n) — the ***sum of all frequencies***:

n =

There are variation series:

– according to the ***frequency variance:***

1) *simple variational series*, in which each variable is met

_____

_____ (p 1);

2) *weighted variational series*, in which variable is met

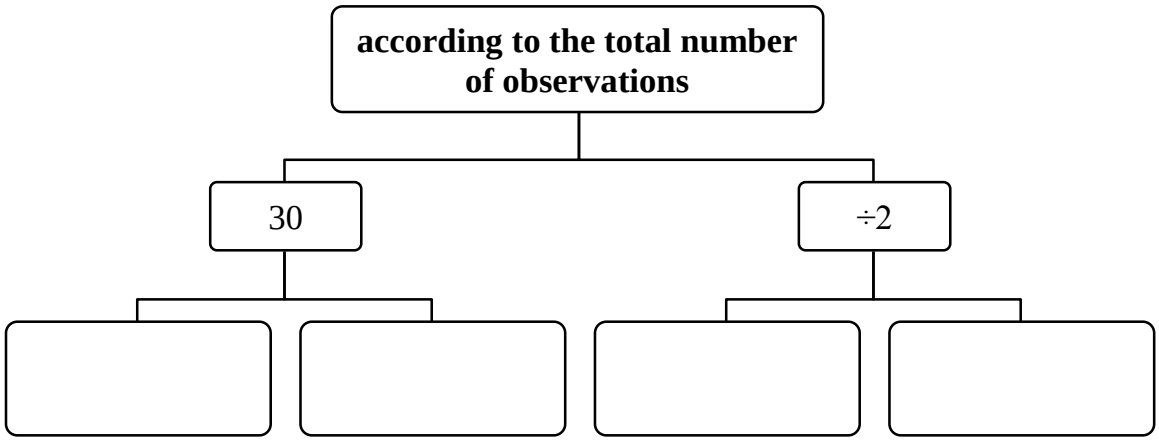
_____

_____ (p 1).

– according to the ***numeric*** value variable

a) _____ *variational series*, in which the variable are integer numbers, such as the number of pulse beats, the number of breaths per minute, the number of days of treatment;

b) _____ *variational series*, in which the variable are fractional numbers, such as weight, height.



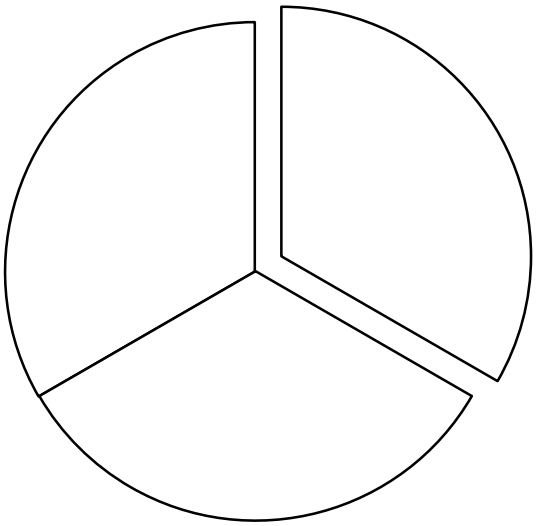
Specify the classification of variation series.

Is the image in the photo a variation series?

If yes, describe this series:

_____
_____
_____
_____
_____

Types of Averages
(Fill in the gaps with the names of typical average values)



_____ is the value of the sign that is most often met in population. It is the value of the variate for which the frequency is maximum.

_____ is the value of the sign that is situated in the middle of the variational series. It divides the variational series into two equal parts.

The **arithmetic mean** (M) is calculated on the basis of all variable of a studied sign. There are simple and weighted arithmetic mean.

The equation for calculating the simple arithmetical mean is:

$$M = \frac{\quad}{n}$$

The equation for calculating the weighed arithmetical mean is:

$$M = \frac{\quad}{n}$$

The methodology and procedure for calculating the arithmetic mean (steps):

1. _____
2. _____
3. _____

The place of the median:

• for even variational series — the place of median = ;

• for odd variational series — the place of median = 2.

TASKS

PART A

Task 1

An analysis of hospital care to patients with acute pancreatitis in hospital A. The following data were obtained.

The length of hospital stay, number of days in a hospital	The number of patients, people
14	2
15	6
16	12
18	10
21	5
TOTAL:	35

1. Describe the variation series (classification).
2. Calculate variation parameters: arithmetic mean (M); mode; median.

Task 2

A group of hypertensive patients with the following parameters in blood pressure were recruited for a study on the treatment of hypertension.

Systolic blood pressure, mm Hg	The number of patients, people
160	4
165	6
170	20
175	12
180	5
TOTAL:	47

1. Describe the variation series (classification).
2. Calculate variation parameters: arithmetic mean; mode; median.

Task 3

As a result of a study on the state of the cardiovascular system in skiers, the following data were obtained.

Heart rate, min-1	Number of study, people
52	3
54	5
56	16
58	10
60	6
TOTAL	40

1. Describe the variation series (classification).
2. Calculate variation series parameters: arithmetic mean; mode; median.

PART B

Task 1

A selective study was conducted, the body length of newborn girls (cm): 49, 48, 50, 55, 52, 50, 49, 53, 51, 51, 50, 52, 52, 53, 54, 49, 50, 51, 52, 53, 52, 50, 51, 50, 51, 53, 52, 49, 53, 51, 51, 50, 52, 52, 53.

1. Build a variation series.
2. Describe the variation series (classification).
3. Calculate the variation series parameters: arithmetic mean, mode, median.

Task 2

A selective study was conducted, where the level of hemoglobin in patients with helminthiasis (g / l) was: 81, 93, 89, 95, 93, 92, 91, 96, 95, 94, 89, 96, 94, 92, 90, 84, 87, 91, 88, 89, 100, 92, 88, 83, 84, 91, 92, 93, 93, 93, 89, 93, 92, 91, 86.

1. Build a variation series.
2. Describe the variation series (classification).
3. Calculate the variation series parameters: arithmetic mean, mode, median.

Task 3

A research on the length of hospital treatment of patients with obstructive bronchitis was performed in the hospital. These are the results (days): 12, 13, 14, 10, 16, 18, 18, 12, 18, 12, 11, 18, 21, 13, 18, 16, 11, 17, 16, 15, 10, 16, 15, 14, 17, 20, 17, 24, 12, 12, 16, 15, 16, 16, 14.

1. Build a variation series.
2. Describe the variation series (classification).
3. Calculate the variation series parameters: arithmetic mean, mode, median.

Task 4

A research on the length of hospital treatment of patients with acute pneumonia was performed in the hospital. These are the results (days): 25, 12, 13, 24, 23, 21, 22, 21, 18, 18, 14, 14, 15, 16, 16, 20, 29, 18, 18, 17, 17, 19, 26, 27, 21, 19, 21, 20, 17, 15, 24, 23, 16, 18, 16, 18, 18, 20, 29, 18, 18, 17, 17, 19, 26, 27, 21, 17, 21, 20, 15, 24, 23, 16, 18.

1. Build a variation series.
2. Describe the variation series (classification).
3. Calculate the variation series parameters: arithmetic mean, mode, median.

CHARACTERISTIC DISTRIBUTION OF VARIABLES IN THE SELECTIVE TOTALITY

_____ is the property of sample that allows equating distribution of studied sign in a sample and distribution of this sign in the population with certain error probability.

The factors influenced on the representativeness are:

– _____ *factors*. There are sampling errors. These errors are amenable to formal calculation;

– _____ *factors* (for example, conformity of the selection method with the goals of research, the correct method of collecting information).

The distribution can be _____ and **discrete**.

The most useful distribution of continuous distributions is **normal distribution.** Normal distribution has an important feature _____ _____ _____.

This distribution is characterized by the coincidence of arithmetic mean value, mode and median also (M Mo Me). The graph of the normal distribution has the form of the _____ curve (Figure 1).

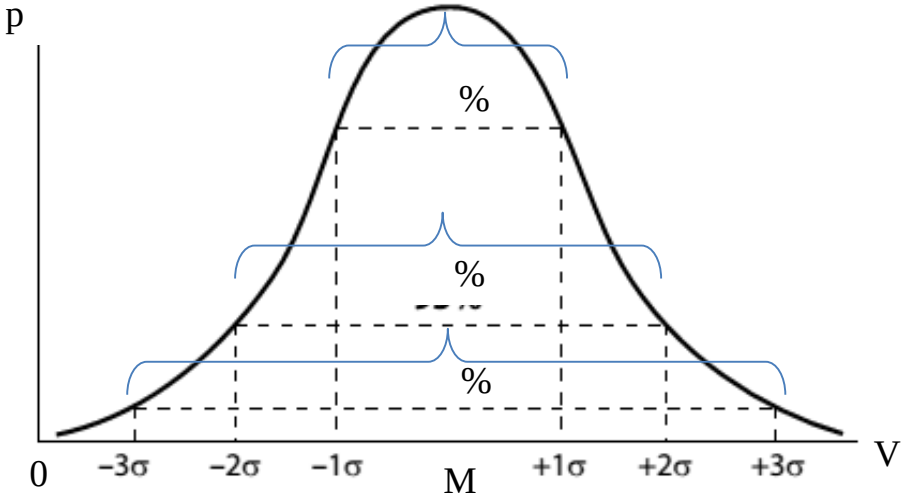


Fig. 1. Connection standard deviation with the structure of the variational series

The most important indicators that are characterized by the distribution of the studied sign are:

1.

2.

3.

Characteristics of the variation series is standard deviation (σ), which shows scattering of the studied signs concerning arithmetic mean. It can be determined by the formula:

1) **n > 30**

$$\sigma = \pm\sqrt{\sum}$$

2) **n ≤ 30**

$$\sigma = \pm\sqrt{\sum(d^2 \times p)}$$

According to a normal law of errors distribution (discovered by K. Gauss and P. Laplas) individual values of the sign are within limits of M ± 3σ, which encompasses 99.7 % of all totality units.

If M ± 2σ, then in the limits of the obtained values there are 95.5 % of all members of the variation series, and finally if M ± 1σ, then in the limits of the obtained values there will be 68.3 % of all members of the variation series.

If the standard deviation is constant and the value of the arithmetic mean is changing, the shape of the normal curve remains unchanged and only its graph moves to the right or to the left on the abscissa axis (Figure 2).

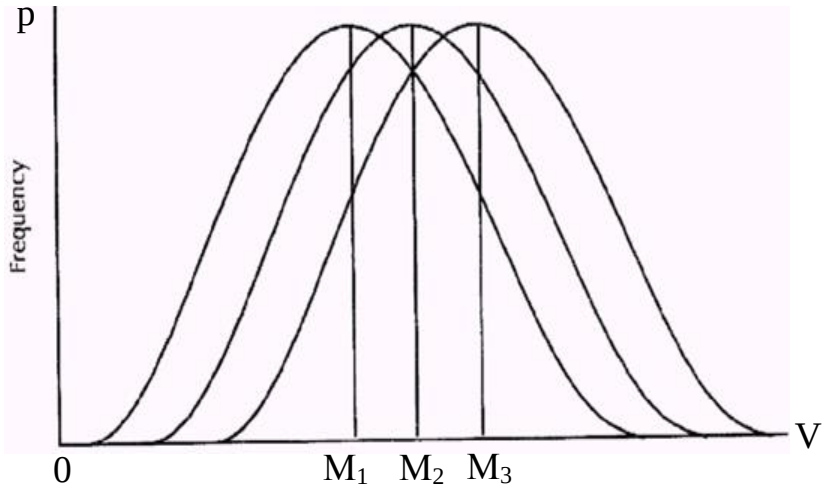


Fig. 2. The identical shape of curves differ by the arithmetic mean (M_1 M_2 M_3; σ_1 σ_2 σ_3)

Arrange the characters in a proper way: more, less, equal

If the arithmetic mean is constant and standard deviation is changing, that involves a change in the width of the curve only (Figure 3).

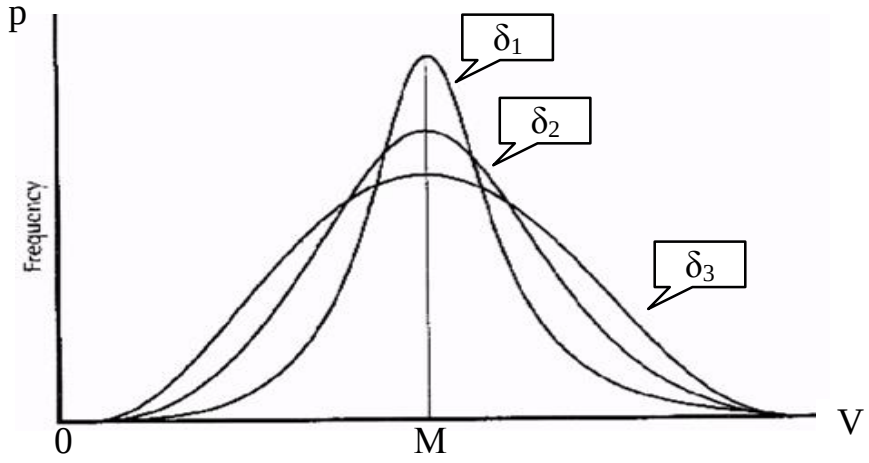


Fig. 3. Three curves where the arithmetic mean is the same (M_1 M_2 M_3) but the standard deviations are different (δ_1 δ_2 δ_3)

Arrange the characters in a proper way: more, less, equal.

In medicine the concept of norm is associated with M ± σ.

In biomedical statistics the rule of three sigmas is used in studying physical development, estimation of health care establishments activity, estimation of population health.

Thus, standard deviation serves for characteristics of signs variety degree which are determined by coefficient of variation:

$$C_v = \frac{\sigma}{M} \times 100\ (\%)$$

Coefficient of variation:

– less than _____ % means weak variety of signs;

– from _____ % to _____ % — middle;

– more than _____ % — strong variety.

To a certain extent coefficient of variation is a criterion of arithmetical mean reliability.

Statistical methods are categorised based on trait distribution (Fill the gaps).

_____	_____
Give examples of methods. _____ _____ _____	***Give examples of methods.*** _____ _____ _____

TASKS

PART A

Task 1

A study on treatment duration at the cardiology department was conducted. A variation series was constructed (the results are presented in the Table 1).

Table 1

Calculation of the arithmetic mean duration of treatment

Duration of treatment (in days), V	Number of patients, p	V × p
15	2	30
16	3	48
17	4	68
18	2	36
19	2	38
21	1	21
22	2	44
Total	*16*	*∑285*

The variation series is weighted, discontinuous, even, small.

$$M = \frac{\sum(V \times p)}{n} = \frac{285}{16} = 18.8 \text{ days}$$

Calculate *the standard deviation, coefficient of variation, estimate the coefficient of variation.*

Task 2

Doctors conducted a study on chest circumference (cm) in 2-month-old girls. The results are shown in the Table 2 as a variation series.

Table 2

Calculation of the arithmetic mean the circumference of the chest (cm) in girls aged 2 months

Circumference of the chest (cm), V	Number of patients, p	V×p
34.6	1	34.6
35.0	1	35.0
35.4	1	35.4
36.0	2	72.0
36.7	1	36.7
37.0	2	74.0
37.6	4	150.2
38.0	5	189.9
38.8	3	116.4
39.2	2	78.5
39.7	3	119.0
40.0	3	120.0
40.9	2	81.8
Total	*30*	*1143.5*

The variation series is weighted, discontinuous, even, small.

$$M = \frac{\sum(V \times p)}{n} = \frac{1143.5}{30} = 38.1 \text{ cm}$$

Calculate *the standard deviation, coefficient of variation, estimate the coefficient of variation.*

PART B

Task 1

Mo = Me = M

Is it typical for a normal distribution?

Task 2

Mo > Me > M

Is it typical for a normal distribution?

PART C

Which samples are clearly not representative and why?

Task 1

The research aims to study the physical development of teenagers in the city of Minsk:

a) young people who are under the supervision of children's clinics № 1, № 5 and № 16 in Minsk;

b) young people studying in the Ballet School;

c) teenagers from secondary schools in the city of Minsk.

g) Suvorov Military School cadets.

Task 2

The aim of the research is to study the physical development of students at the BSMU.

a) The first-year students from all faculties;

b) all the students of pediatric faculty;

c) all the students of medical, pediatric, dental and medical-prophylactic faculties;

g) every third student from each group in the BSMU.

PART D

Task 1

A sample study was conducted to study the duration of hospital treatment among the patients with obstructive bronchitis. The following results were obtained (in days): 12, 13, 14, 10, 16, 18, 18, 12, 18, 12, 11, 18, 21, 13, 18, 16, 11, 17, 16, 15, 10, 16, 15, 14, 17, 20, 17, 24, 12, 12, 16, 15, 16, 16, 14.

Based on the above data:

1. Make the variation series and determine its type.
2. Determine the mode and median.
3. Calculate:
   a) the arithmetic mean;
   b) standard deviation;
   c) coefficient of variation, estimate the coefficient of variation.

Task 2

There was a sample study on the duration of treatment in children with acute pneumonia. The following results were obtained (in days): 25, 12, 13, 24, 23, 21, 22, 21, 14, 14, 15, 16, 16, 20, 29, 18, 18, 17, 17, 19, 26, 27, 19, 21, 21, 20, 17, 15, 24, 23, 16, 18, 18, 18, 20

Based on the above data:

1. Make the variation series and to determine its type.

2. Determine the mode and median.
3. Calculate:
   a) the arithmetic mean;
   b) standard deviation;
   c) coefficient of variation, estimate the coefficient of variation.

RELIABILITY ESTIMATION OF THE RESULTS OF STATISTIC INVESTIGATION

Random errors of representativeness (_____) — the actual difference between the average and relative values obtained during the sampling and similar values that would be obtained in a study on the entire population.

_____ **—** bounds which with a certain probability faultless prognosis entered actual value (average or relative) of statistic attribute characterizing the entire population.

Estimation of reliability of the difference between the average values of the two samples

When selecting units of observation some errors are possible. These errors are objective and natural. When determining the degree of accuracy of sampling, a possibility of error value is evaluated.

Such errors are called _____ _____ **(m)**. It is the actual difference between the average and relative values obtained during the sampling and similar values that would be obtained in a study on the general population.

In practice, to determine the average sampling error in statistical research, the following formulas are used.

Calculation error of representativeness average value

Calculation error of representativeness (m_M) arithmetic mean value (M) used following formula:

$$m_M = \pm \frac{\sigma}{\sqrt{n}},$$

where σ — _____;
n — _____

When the number of observations is less than 30, following formula should be used:

$$m_M = \pm \frac{\sigma}{\sqrt{n-1}},$$

where σ — _____;
n — _____

Calculation error of representativeness the relative value

In order to calculate a standard error (m_p) of the relative value (P) the following formula is used:

$$m_p = \pm \sqrt{\quad},$$

where P — the corresponding relative value; q = 100 – P, if relative value calculated in per cent ( ); q = 1000 – P, if relative value calculated in ppm ( ); q = 100000 – P, if relative value calculated in prosantimille ( ); n — _____

When the number of observations is less than 30, the following formula should be used:

$$m_p = \pm \sqrt{\quad},$$

where P — the corresponding relative value; q = 100 – P, if relative value calculated in per cent (%); n — _____

Remember, the rate should be 3 times more than their error of representativeness:

$$\frac{P_{sample}}{m_p} \geq 3,$$ therefore the sample is representative.

To determine the accuracy with which the researcher wants to get the result the **probability of faultless prognosis is used**. It is characteristic of the reliability of the results of sampling of biomedical statistical research. Usually, using probability of faultless prognosis is ____ **% or** ____ **%** in the statistical medical and biological research.

Certain degree of probability of faultless prognosis corresponds to a certain _____ **error** of random sampling ( ). This value is determined by the formula:

$$= t \times m,$$

where t — Confidence coefficient; m — Standard error (SE).

Calculation confidence interval

For determining frames of confidence following formula should be used:

– for the average mean

$$\underline{M} = \quad ' \quad = M' \pm t \times m_M,$$

where $\underline{M}$ — the confidence interval of the average value in general population; M' — the average value obtained in the research on sample population; t — confidence coefficient whose value is determined by the degree of probability of faultless prognosis with which the researcher wishes to obtain a result; m_M —standard error of average.

– the relative values

$$\underline{P} = P' \pm \Delta = \quad ' \quad ,$$

where $\underline{P}$ — the confidence interval of the relative values in the general population; P' — the relative values obtained in the research on sample population; _____ —

confidence coefficient whose value is determined by the degree of probability of faultless prognosis with which the researcher wishes to obtain a result; _____ — standard error of relative values.

$$df = n' = n - 1,$$

where n — the number of sample.

Table 1

Table of values t (Student criterion)

The degrees of freedom n'	**Probability of error**		
	0.05 = 5 %	0.01 = 1 %	0.001 = 0.1 %
1	12.70	63.66	637.59
2	4.30	9.92	31.60
3	3.18	5.84	12.94
4	2.78	4.60	8.61
5	2.57	4.03	6.86
6	2.42	3.71	5.96
7	2.36	3.50	5.31
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
11	2.20	3.11	4.44
12	2.18	3.06	4.32
13	2.16	3.01	4.22
14	2.14	2.98	4.14
15	2.13	2.95	4.07
16	2.12	2.92	4.02
17	2.11	2.90	3.96
18	2.10	2.88	3.92
19	2.09	2.86	3.88
20	2.09	2.84	3.85
21	2.08	2.83	3.82
22	2.07	2.82	3.79
23	2.07	2.81	3.77
24	2.06	2.80	3.75
25	2.06	2.79	3.73
26	2.06	2.78	3.71
27	2.05	2.77	3.69
28	2.05	2.76	3.67
29	2.04	2.76	3.66
30	2.04	2.75	3.64
∞	1.96	2.58	3.29

TASKS

Task 1

1230 students have to undergo a medical examination. 1200 students were examined. 455 students were diagnosed with scoliosis (curvature of the spine).

Calculate:

– point-prevalence (intensive index per 1000 examined patients);

– error of representativeness and assess the reliability of sampling;

– frames of confidence used probability of faultless prognosis 95.0 %;

– frames of confidence used probability of faultless prognosis 99.9 %.

Task 2

During medical examination of 200 workers which had violated the common diet, liver disease and biliary tract were found in 40 people.

Calculate:

– point-prevalence (intensive index per 1000 examined patients);

– error of representativeness and assess the reliability of sampling;

– frames of confidence used probability of faultless prognosis 95.0 %;

– frames of confidence used probability of faultless prognosis 99.9 %.

ESTIMATION OF RELIABILITY DIFFERENCE OF STATISTICAL VALUES

When conducting a medical and biological research on two compared samples there is a need to determine not only the difference between them, but its reliability too. To assess the reliability of differences of the compared values the following formula is used:

– *the* _____

$$t = \frac{M_1 - M_2}{\_\_\_\_\_}$$

– *the* _____ ***values***

$$t = \frac{P_1 - P_2}{\_\_\_\_\_},$$

where M_1, M_2, P_1 and P_2 — statistical values obtained during sample surveys; m_1 and m_2 — the representativeness of their mistakes; t — the coefficient of reliability.

Differences are significant for **t >** _____ (the number of observations is greater than 30); which corresponds to the probability of faultless prognosis equal to or greater than 95 %.

TASKS

Task 1

Evaluate the reliability of the differences between t**he average bed occupancy rate in the town children's hospital (264.2 ± 6.2) and the regional children's hospital (322.8 ± 4.4).** Determine if those rates are significantly different. Calculate the confidence intervals for both average bed occupancy rates at a significant level of p=0.05.

Task 2

Assess the reliability of differences in hepatitis morbidity between two cities: City A with 45 cases out of 12,000 people, and City B with 1,500 cases out of 450,000 people. **Are these indicators significantly different? Calculate the confidence interval for both morbidity rates (p = 0.05).**

PAIRED t-TEST ASSUMPTIONS

For the ***paired t-test***, we need two variables. One variable defines the pairs for the observations. The second variable is a measurement. Sometimes, we already have the paired differences for the measurement variable. Other times, we have separate variables for "before" and "_____" measurements for each pair and need to calculate the differences.

To apply the paired t-test to test for differences between paired measurements, the following assumptions need to hold:

Subjects must be independent. Measurements for one subject do not affect measurements for any other subject.

Each of the paired measurements must be obtained from the same subject. For example, the before-and-after weight for a patient must be from the same person.

The measured differences are _____ distributed.

The formula of the paired t-test is defined as the sum of the differences of each pair divided by the square root of n times the sum of the differences squared minus the sum of the squared differences, overall n – 1.

The formula of the paired t-test:

$$t = \frac{M_d}{\sigma_d / \sqrt{n}}$$

Fill the gaps.

t _____

M_d _____

σ_d _____

n_____

TASKS

PART A

To evaluate the effectiveness of a new anti-anemia agent, hemoglobin levels in patients with anemia were measured before and after administering the drug. The **paired t-test** will be used to determine if there are statistically significant differences in hemoglobin levels before and after taking the medication.

Task A-1

The mean difference in hemoglobin levels measured before and after taking the drug $M_d = 22.0$ g/L, standard deviation of differences $\sigma_d = 12.5$ g/L, number of patients $n = 12$.

Task A-2

The mean difference in hemoglobin levels measured before and after taking the drug $M_d = 20.0$ g/L, standard deviation of differences $\sigma_d = 15.7$ g/L, number of patients $n = 10$.

Task A-3

The mean difference in hemoglobin levels measured before and after taking the drug $M^d = 21.0$ g/L, standard deviation of differences $\sigma^d = 15.9$ g/L, number of patients $n = 14$.

Task A-4

The mean difference in hemoglobin levels measured before and after taking the drug $M^d = 22.0$ g/L, standard deviation of differences $\sigma^d = 13.1$ g/L, number of patients $n = 10$.

PART B

Comparison of peak expiratory flow rate (PEFR) before and after a walk in a cold winter's day for a random sample of 9 asthmatics. The table contains two columns: one for PEFR before the walk and the other for PEFR after the walk. Each row represents the same subject.

Subject	Before	After
1	312	300
2	242	201
3	340	232
4	388	312
5	296	220
6	254	256
7	391	328
8	402	330
9	290	231

CORRELATION

Correlation — _____
_____

Functional relationship — relationship where each value of one of attributes corresponds to _____
_____

Correlation relationship — relationship where each value of one of attributes corresponds to _____
_____

Rankings — _____
_____

Correlation

Types of manifestations of quantitative relationships between attributes:
– Functional relationship;
– Correlation.

Functional relationship (give examples) _____
_____

Correlation _____
_____

Ways of presenting correlation:

1.

2.

The correlation coefficient indicates the direction and strength of the relationship between attributes. The limits of its variations are from 0 to ± ___.

Methods for determining the correlation coefficient:
– _____ method (Pearson);
– _____ method (Spearman).

When both variables are normally distributed it is better to use Pearson's correlation coefficient (parametric method), otherwise Spearman's correlation coefficient (_____ method) is used.

Direction of correlation:
– _____
– _____

When one variable is increasing together with the other one the correlation is considered as positive; when one is decreasing while the other is increasing it is negative.

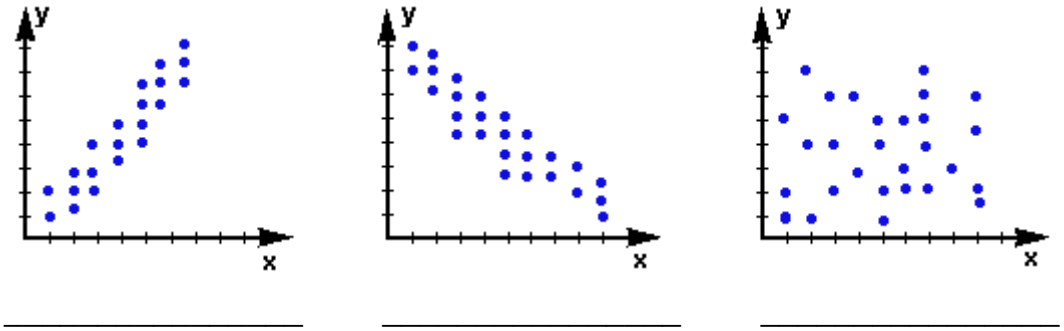
The strength of the correlation (fill the table).

The strength	Positive	Negative
Strong		
Average		
	0.001–0.299	**(–0.299) – (–0.001)**

Complete absence of correlation is represented by 0.

If correlation coefficient is _____, it is a functional relationship (absolution).

Type of correlations



_____ _____ _____

Pearson Correlation Coefficient

Pearson correlation coefficient may be used if:

1.

2.

3.

The methodology and procedure for calculating the Pearson correlation coefficient:

1. Construct a variation series for each of the compared attributes, denoting the first and second series of numbers x and y respectively;

2. Determine for each variation series _____ (and);

3. Find the _____ (d_x and d_y) between the numerical value and the mean value variants of the its variation series ($d_x = V - M_x$; $d_y = V - M_y$);

4. Multiply the received deviation ($d_x \times d_y$);

5. Square each deviation and sum up each series ($\sum d^2_x$ and $\sum d^2_y$);

6. Substitute the values into the formula for calculating the correlation coefficient:

$$r_{xy} = \frac{\sum \quad}{\quad}$$

7. Assess the value of the correlation coefficient r.

Calculation correlation coefficient error

The formula for calculating the Pearson correlation coefficient error is:

$m_{rxy} = \frac{1 - r^2 xy}{\sqrt{n-2}}$, when the number of observations is less than _____;

$m_{rxy} = \frac{1 - r^2 xy}{\sqrt{n-1}}$, when the number of observations is from _____ to _____;

$m_{rxy} = \frac{1 - r^2 xy}{\sqrt{n}}$, when the number of observations is more than _____.

Evaluation of the reliability of the correlation coefficient

Reliability of the Pearson correlation coefficient is determined by the formula:

$$t = \frac{}{m_{rxy}}$$

Criterion **t** is measured at the table of values **t**, considering the number of degrees of freedom (n – 2), where **n** is a number of paired variants. Criterion t must be **equal to or greater** than the table corresponding to the probability of faultless prognosis ≥ _____ %.

Spearman's rank correlation

Spearman's rank correlation may be used if:

– Non-parametric measure of correlation — the data of the variables not to be _____ distributed in the population.

– Its special case of Pearson's correlation coefficient in which the two sets of data are converted to _____.

The methodology and procedure for calculating the Spearman's rank correlation:

1. Determine the ranks by value of each value of the series. If the first series (X) ranges from lower to greater value, the second series (Y) should be ranked in the same way.

2. Determine the _____ between the ranks of each pair of series (X) and series (Y): d_{xy} = (X) – (Y). They are equal to zero taking into account the sum of signs.

3. Square the received difference and sum them up (Σd^2).

4. Calculate the coefficient of rank correlation:

$$\rho = 1 - \frac{}{n\,(n^2 - 1)}$$

5. Assess the value of the correlation coefficient **ρ**.

Calculation Spearman's rank correlation error

The formula for calculating the Spearman's rank correlation error is:

$$m_{\rho xy} = \sqrt{\frac{}{n-2}}$$

Evaluation of the reliability of the correlation coefficient

Reliability of the Spearman's rank correlation coefficient is determined by the formula:

$$t = \frac{\rho xy}{}$$

Criterion **t** is measured at the table of values **t**, considering the number of degrees of freedom (n – 2), where **n** is number of paired variants. Criterion t must be **equal to or greater** than the table corresponding to the probability of faultless prognosis ≥ _____ %.

TASKS

Pearson Correlation Coefficient

Task 1

Calculate the correlation coefficient to determine the direction and strength of the connection between the amount of calcium in water and the water hardness, if following data were obtained.

Hardness of water (degrees)	The amount of calcium in water (mg/l)
4	28
8	56
11	77
27	191
34	241
37	262

Assess its reliability.

Task 2

Calculate the correlation coefficient to determine the direction and power of the connection between age and frequency of disability, if following data were obtained.

Age (years)	Disability (per 10 000 population)
25	0.3
35	0.6
45	0.7
50	1.8
55	4.1
60	4.2

Assess its reliability.

Task 3

The pulse of 11 patients was measured at rest (before physical activity) and after 15 squats. Determine whether there is a correlation between the pulse before and after physical activity.

If there is, describe it.

Pulse before exercise (X)	Pulse after exercise (Y)
74.00	95.00
74.00	95.00
73.00	98.00
72.00	92.00
72.00	97.00
75.00	104.00
75.00	91.00
74.00	96.00
74.00	98.00
69.00	95.00
77.00	99.00

Assess its reliability.

Spearman's rank correlation

Task 1

Calculate the rank correlation coefficient to determine the direction and strength of the connection between age and incidence of scarlet fever, if following data were obtained.

Age (years)	Incidence of scarlet fever (100 000 population)
15	199
25	178
35	129
45	16
55	24
65	3

Assess its reliability.

Task 2

Calculate the rank correlation coefficient to determine the direction and power of the connection between age and incidences of breast cancer in women, if the following data were obtained.

Women Age (years)	Incidence of breast cancer (100 000 population)
15	1
25	16
35	14
45	56
55	65
65	123

Assess its reliability.

REGRESSION ANALYSIS IN MEDICINE

In order to establish how it can change one unit when changing the other one, it is necessary to use regression analysis (method of regression).

Regression — _____
_____.

The regression coefficient is an absolute amount by which the average changes the second feature when changing the linked first unit.

As changeable values (x and y) and the regression are interrelated, to calculate two regression coefficients R_{xy} and R_{yx} the following formula should be used:

R_{xy} =

R_{yx} =

Where *r* is the correlation coefficient; σ_x and σ_y — standard deviation of the two compared rows.

The regression coefficient describes only linear dependence.

We can calculate the margin error (m_R) for the regression coefficient (R_{xy}). Error of the regression coefficient is equal to the error of the correlation coefficient (m_r), multiplied by the ratio of the standard deviations:

$$m_R = \frac{\sigma_x}{\sigma_y} \times$$

The criterion of reliability of the regression coefficient is calculated by the formula:

$$t_R = \text{———}.$$

The accuracy of the value t is a table coefficient with $n' = n - 2$, where n is the number of pairs of observations.

Using the regression coefficient without special measurements, one can determine the value of one of the signs (e.g., body weight), knowing the value of another (growth). For this purpose the linear regression equation.

$$y = M_{...} + R_{xy} \times (x - M_{...}),$$

where y — the desired value of the mass of the body; M_y — mean value of body mass that is typical for this age; R_{xy} — the regression coefficient of body mass growth; x — value of growth; M_x — mean of growth.

Nonlinear regression is a method of finding a nonlinear model of the relationship between the dependent variables and set of independent variables.

The individual values of individual characteristics are very diverse. For example, in a wide range can range indicators of body weight and chest circumference in people with the same height. The measure of diversity individual size of the signs characterizes Sigma regression $\sigma_{Ry/x}$.

$$\sigma_{Ry/x} = \sigma_{...}\sqrt{1 - \quad},$$

where _____ — Sigma (standard deviation) of the regression; _____ — standard deviation of trait y; _____ is the correlation coefficient between variables x and y.

Regression coefficients, regression equations and Sigma of the regression is widely used for making regression scales that are used in individual estimation of physical development.

Scale regression (***fill the gaps***).

Growth, cm	Average weight, kg	The smallest weight, kg	The biggest weight, kg
x_1	y_1	$y_1 \ldots \sigma_{Ry/x}$	$y_1 \ldots \sigma_{Ry/x}$
x_2	y_2	$y_2 \ldots \sigma_{Ry/x}$	$y_2 \ldots \sigma_{Ry/x}$
x_3	y_3	$y_3 \ldots \sigma_{Ry/x}$	$y_3 \ldots \sigma_{Ry/x}$

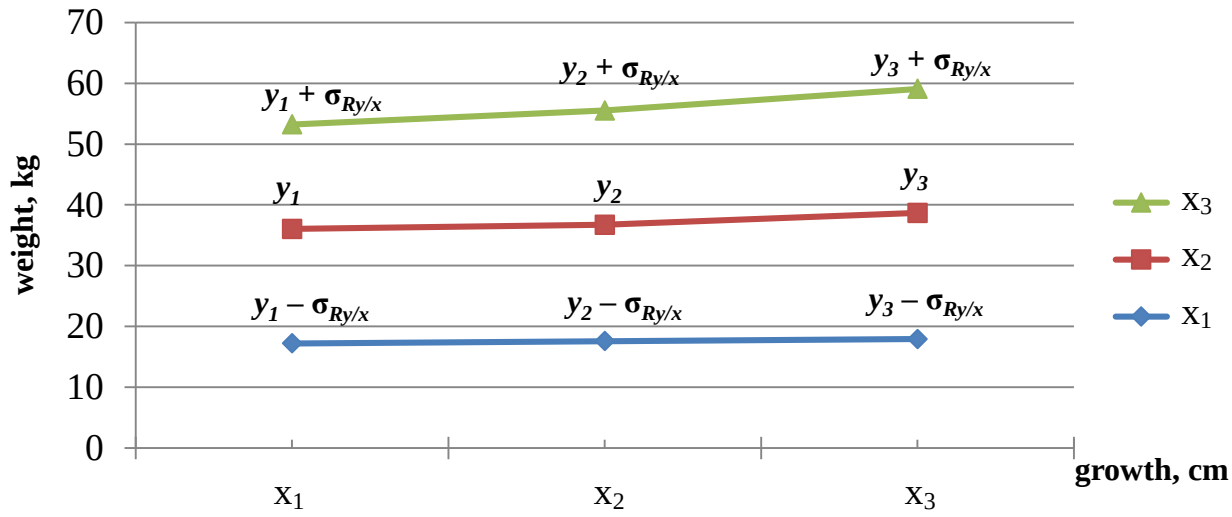


Fig. 1. The scale of the regression of body mass and growth

The correlation coefficient (r) has no meaningful interpretation. However, its square is called the coefficient of determination (r^2).

The coefficient of determination r^2 is a measure of how changes in the dependent characteristic are influenced by changes in the independent. More precisely, it is the proportion of the variance independent of the characteristic, explained by the influence of the dependent.

$$\_\_\_\_\_ = 1 - \frac{\sigma_x^2}{\sigma_y^2},$$

where _____ — coefficient of determination.

If two variables are functionally linearly dependent, then we can say that the change of the variable y is fully explained by the change in the variable x, and this is precisely the case when the coefficient of determination equal to one , $r^2 = 1$ (the correlation coefficient can be equal to 1 and –1).

_____ *interval for regression coefficient* (R_{xy}):

$$(R_{xy} - t \times m_R, R_{xy} + t \times m_R),$$

where R_{xy} — regression coefficient; m_R — error of regression coefficient; t — factor, the value of which is determined by the degree of probability of the faultless prognosis with which the researcher wishes to obtain the result.

TASKS

Task 1

The dependence between the frequency of incidence of bronchitis and duration of smoking.

Duration of smoking, years	3	4	5	6	7	8	9	10
Frequency of bronchitis, %	6	9	12	13	14	21	26	35

Make a hypothesis of the study to calculate the correlation coefficient, assess its validity, calculate the regression coefficient, draw the regression line. Make a conclusion.

Task 2

The dependence of the number of acute respiratory infections (ARI) in the territory and air temperature in winter from December to February.

Number of ARI	30	31	33	34	34	36	38	39	38	36	28	34
-t °C air	20	21	22	23	21	25	25	29	28	23	20	22

Make a hypothesis of the study to calculate the correlation coefficient, assess its validity, calculate the regression coefficient, draw the regression line. Make a conclusion.

Task 3

The incidence of myocardial infarction in men and women of different age in the reporting year (100 000 population).

Gender	Age (years)					
	40	45	50	55	60	65
Men	35	135	160	220	650	700
Women	12	18	35	100	130	350

Make a hypothesis of the study to calculate the correlation coefficient between age and morbidity in both men and women, to evaluate its accuracy, calculate the regression coefficient of incidence by age, determine the expected incidence for men and women at the age of 40, 50, 70 years , draw the regression line on the found points. Make a conclusion.

TIME SERIES

While studying the _____ of any phenomenon, researchers often use a method of time series.

_____ **series** is a series of equal statistical values showing the change of any phenomenon in time and in chronological order after a certain period of time.

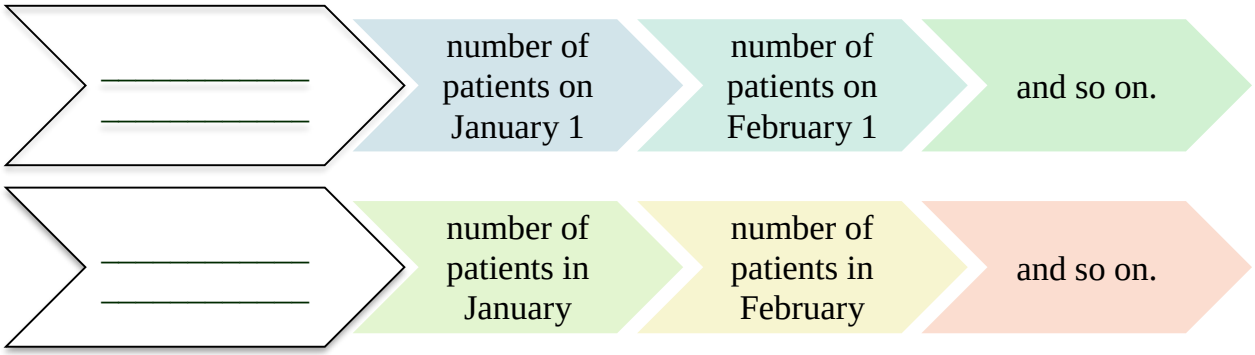
The numbers that make up the dynamic range, called _____.

The level number is the size (value) of certain phenomena, made in a certain period or to a specific point in time.

The levels of a number can be represented in absolute, relative and average values.

Series are divided into _____ (consisting of absolute values) and _____, consisting of a relative or averages.

The time series can be momentary and interval (*fill the gaps*).



Methods of alignment of time series

There can be used following techniques:

_____ is produced by summing up the data for the number of periods. As a result we are getting longer periods of time.

_____ is determining the average value of each consolidation period. For this purpose it is necessary to summarize the levels of the neighboring periods, and then the sum is divided by the number of summands. This provides greater clarity changes over time.

_____ to eliminate the effect of random fluctuations in the levels of time series and clearly reflects the trend of the phenomenon. Most often, we sum up three members of the series, but you can take more.

Indicators of time series

First of all, the series can be characterized by the values of the members of the series, called equations. The value of the first member of the series is called the initial level, the value of the last member of the series — the final level. The average value of all members is called the secondary level.

_____ — the difference between the current and previous levels; positive difference represent the increase while negative difference represent the decrease. The value of gains or losses reflect changes in the levels of dynamic range for a certain period of time.

_____ — shows the relationship of each subsequent level to the previous level and is usually expressed in percentage.

_____ — the ratio of the absolute gain or loss of each successive member of the series to the previous level, expressed as a percentage. The growth rate can also be calculated by the formula: **Growth rate — 100 %**.

_____ obtained by dividing the absolute value of the increase or decrease in the rate of growth or decline over the same period.

TASKS

Task 1

Make an alignment of cases of dysentery (Table 1).

Table 1

Distribution of data on dysentery incidences by month

Months	The number of patients	… The number of patients for the period of 3 month	Group average	Moving average
1	5			
2	3			
3	4			
4	5			
5	8			
6	9			
7	13			
8	20			
9	14			
10	11			
11	7			
12	4			

Complete the table.
Create a chart "Monthly Dysentery Dynamics in the City N".

Task 2

Using this data, determine the type of Time series, calculate indicators, and analyze the results.

Prevalence of TB in the Republic N's urban population (first-time diagnosed cases per 100,000 people):

1st — 42.7
2nd — 49.6
3rd — 48.8
4th — 51.0
5th — 47.9

Task 3

Determine the type of Time series, calculate indicators, and analyze the data based on these figures.

Tuberculosis prevalence among rural residents in the Republic N (new cases per 100,000 people):

1st — 59.1
2nd — 57.2
3rd — 59.9
4th — 66.6
5th — 70.8

Task 4

Using these data, determine the type of Time series, calculate indicators, and analyze the results.

Chlamydial disease prevalence in the Republic N (number of newly diagnosed cases per 100,000 people):

1st — 122.2
2nd — 155.3
3rd — 177.7
4th — 204.7
5th — 238.4

FORECASTING METHODS

Forecasting — _____
_____

Define the time period for the following forecasting terms:

short-term (up to _____ years), ***medium term*** (_____ years) and ***long term*** (over _____ years) forecasting.

Forecasting methods can be divided into several groups:

1) _____
2) _____
3) _____
4) _____

Methods of forecasting health indicators

Extrapolation is a forecasting method based on the analysis of previous years' data. This technique can be applied with available data over several years, though such forecasts are preliminary. More precise predictions can be achieved by using extrapolation that incorporates information about the levels of the phenomenon and analyzing its development process (according to the dynamic range). The method of extrapolation can be used to predict incidence rates between two points.

The formula below predicts disease levels at two points:

$$\mathbf{P}_{...} = \mathbf{P}_{...} + \mathbf{T} \times$$

where $\mathbf{P_t}$ is the predicted level; $\mathbf{P_1}$ — index of morbidity in one of the previous years (closer to target); $\mathbf{P_0}$ is the incidence rate for another year prior to that (more distant from the target); **n** — the period between the two studies (two years); ***T*** — the period between the receipt of the last result and the year of prediction (in years).

Application of the method of extrapolation in forecasting

Forecasting of incidence rates according to dynamic range is carried out after the last alignment and identifies trends of the phenomenon under study.

_____ trends, depending on the nature of the determining causes may be straight or curved. Straight-line trend indicates a uniform change in the intensity of the phenomenon. If the reasons generating the trend phenomena are unevenly distributed and the effect of their actions gradually decreases or increases, the trend becomes curvilinear.

To align the dynamic series of incidence in a straight line use the equation of linear dependence:

$$\mathbf{P_{theory} = A + b * X}$$

where $\mathbf{P_{theory}}$ — _____;
A — _____;
b — _____;
X — _____.

For the synthesis of the quantitative trend assessment of the dynamic series uses the average growth rate expressed in per cent. It characterizes the ***average rate of change*** (***$T_{average\ rate\ of\ change}$***) of incidence rates in dynamic range and is a relative estimate of the rate of change of the level of the series. Expect him on the basis of theoretical values trend line:

$$\mathbf{T_{average\ rate\ of\ change} = (a / b) * 100\ \%}$$

The estimate of the average rate of change of the number is as follows:
– $\mathbf{T_{average\ rate\ of\ change}}$ = 0 to ± 1 %, the trend is _____
– $\mathbf{T_{average\ rate\ of\ change}}$ = ± 1 % to ± 5 % the trend is _____
– $\mathbf{T_{average\ rate\ of\ change}}$ is more than ± 5 % the trend is _____

Confidence limits of the theoretical predicted morbidity are determined then the following formula:

$$\mathbf{P_{for.} = P_{for.\ theor.} \pm t * m}$$

Where Student's _____ — criterion corresponding to the selected level of the faultless prognosis (in biomedical research is usually applied probability of error-free forecast equal to 95.0 % for which t = _____).

m-average error of the predicted theoretical incidence rate: it is calculated according to the formula:

$$\mathbf{m =}$$

where **P** — _____; **q** — the difference between the dimension index and the index (if the rate of incidence calculated per 10,000, q = 10000 – $P_{for.\ theor.}$); **N** _____.

TASKS

Based on the data dynamic range:

Calculate the forecasting incidence for 2026 is two points (2024 and 2022).

Task 1

Morbidity of population of the City N with malignant neoplasms (per 100,000 population): 2017 — 263.7; 2018 — 273.4; 2019 — 279.3; 2020 — 284.3; 2021 — 294.4; 2022 — 296.1; 2023 — 314.9; 2024 — 319.1.

Task 2

Incidence of the adult population of the City N from hypertension (per 100,000 population): 2017 — 303.0; 2018 — 243.0; 2019 — 263.0; 2020 — 335.0; 2021 — 307.0; 2022 — 301.6; 2023 — 359.7; 2024 — 339.0.

Task 3

Morbidity of population of the City N with mental disorders (per 100,000 population): 2017 — 575.9; 2018 — 765.2; 2019 — 837.2; 2020 — 958.5; 2021 — 1061.2; 2022 — 1073.7; 2023 — 1147.3; 2024 — 1152.8.

Task 4

The number of patients with first time diagnosis of drug abuse in the City N (per 100,000 population): 2019 — 1.3; 2020 — 2.8; 2021 — 4.7; 2021 — 4.0; 2022 — 5.2; 2023 — 9.4; 2024 to 9.0.

ANALYSIS OF CONTINGENCY TABLE

Chi-Square-test

The conformity criterion is used for:

1) _____
_____
_____;

2) _____
_____
_____;

3) _____
_____
_____

Features of the methodology for calculating the criterion:

1) only absolute values are used for calculation;

2) confirms the existence of a link, but does not establish its degree;

3) the greater the value of the chi-square is, the more the result differs from the theoretical one.

Null Hypothesis (H_0): _____
_____
_____.

Alternative Hypothesis (H_a): _____
_____
_____.

The chi-square test statistic is calculated by using the following formula:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(P - P_1)^2}{P_1}$$

where **P** is _____;
$\mathbf{P_1}$ is _____

We will compare the value of the test statistic to the critical value of χ^2_α with degree of freedom: **df = (s – 1) (___ – 1)**, where **s** — number of compared groups; … — number of outcomes. We reject the null hypothesis if χ^2 _____ χ^2_α.

Calculation procedure:

1. We present the actual data for the observation groups.
2. We accept the ____ hypothesis.
3. We calculate the "_____" data in accordance with the _____ hypothesis.
4. We calculate the value of the coefficient "_____".
5. _____

TASKS

Task 1

Distribution of tuberculosis patients on the timing of treatment and availability of BC in the sputum

Duration of use of medicines	Outcome		Total
	BC found in the sputum	BC sputum absent	
Up to 6 months	21		27
Over 6 months	9		33

State the null hypothesis. Use the Chi-square hypothesis to test. Calculate the coefficient, evaluate it, make a decision.

Task 2

Mortality among women with renal failure, which underwent an aboriton

Time frame from the date of the operation (in days)	Outcome		Total
	Died	Survived	
1–4	7	44	51
5–22	6	24	30
After 22	16	14	30

State the null hypothesis. Use the Chi-square hypothesis to test. Calculate the coefficient, evaluate it, make a decision.

Task 3

Mortality among patients which were operated on acute appendicitis

The timing of the operation since the onset of the disease (in days)	Outcome		Total
	Died	Survived	
1–2	5		75
3–4	8		36
After 4	3		13

State the null hypothesis. Use the Chi-square hypothesis to test. Calculate the coefficient, evaluate it, make a decision.

CASE-CONTROL STUDIES

Case-control and cohort studies are _____ studies that lie near the middle of the hierarchy of evidence.

Case-control studies are _____. They clearly define two groups at the start: one with the outcome/disease and the other one without the outcome/disease. They look _____ to assess whether there is a statistically significant difference in the rates of exposure to a defined risk factor between the groups.

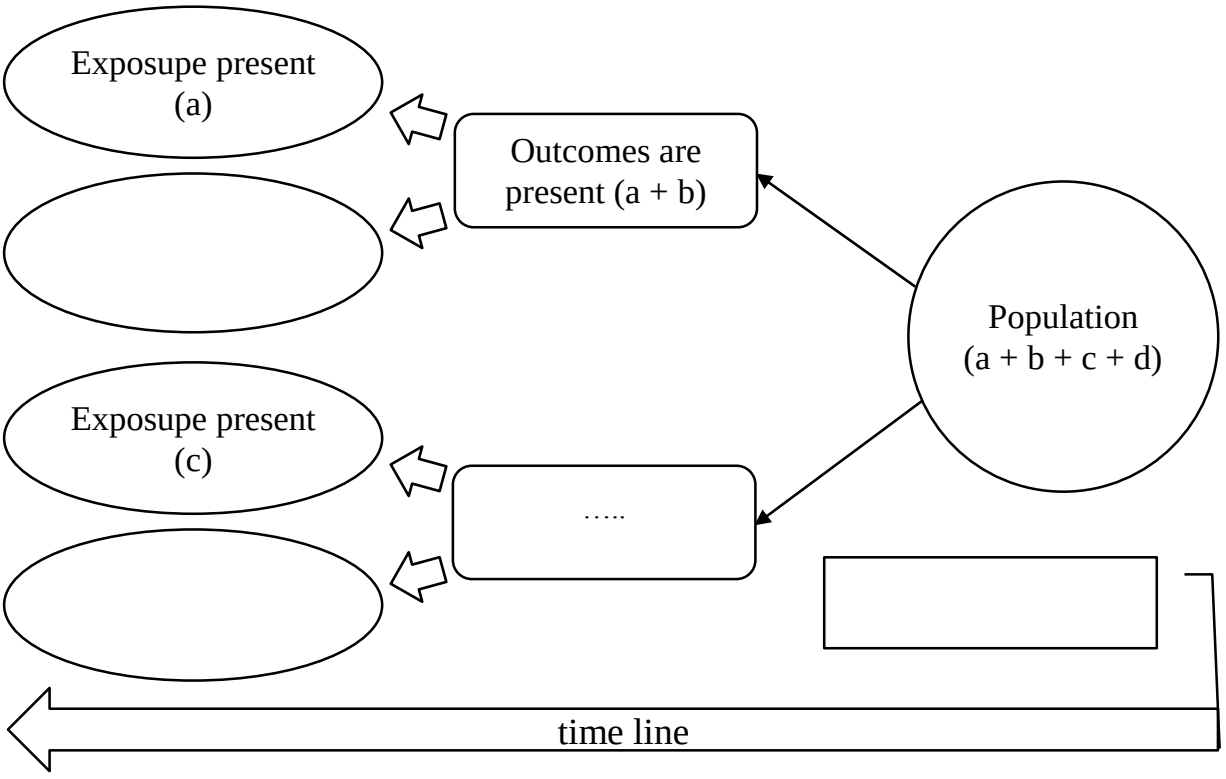


Fig. 1. Case-control study design

The main outcome measure in case-control studies is **odds ratio (OR)**.

Table 1

The results of case-control study

Group	Exposure to risk factor		Total
	…	…	
Outcomes are present	a	b	a + b
…	c	d	c + d
Sum	a + c	b + d	a + b + c + d

OR = (a × d) / (c × b),

where OR — _____; a — _____;
b — _____; c — _____;
d — _____.

Table 2

Advantages and disadvantages of case-control studies (fill the gaps)

……	……
Cheaper	Retrospective / more prone to bias
Quicker / easier to conduct	Can only assess one outcome
Good for diseases with long latency periods	Risk cannot be evaluated
Good for rare diseases	Prevalence cannot be evaluated

COHORT STUDIES

Cohort studies can be retrospective or prospective. Retrospective cohort studies are NOT the _____ as case-control studies.

Prospective cohort studies are more common.

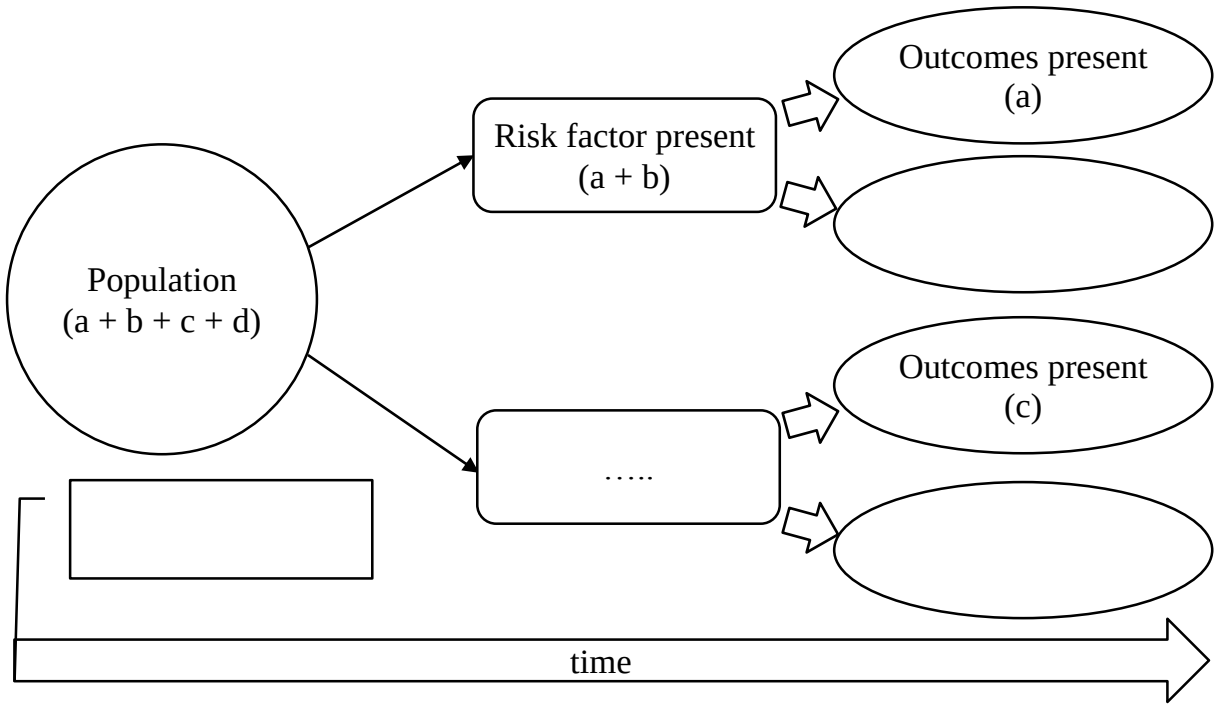


Fig. 1. Cohort study design

Table 1

Advantages and disadvantages of cohort studies

……	……
Prospective	More expensive
Can establish risk directly	Longer / harder to conduct
Can assess multiple outcomes	Cannot be used for rare diseases
Good for rare exposures	Cannot be used for diseases with long latency periods

Cohort studies are good for assessing prognosis, risk factors and harm. The outcome measure in cohort studies is usually a risk ratio / relative risk (RR).

Table 2

2 × 2 table to evaluate the results of cohort studies

Group	**Efficient sign**		**Total**
	+	–	
A risk factor group (Exposed, F+)	a	b	a + b
Control group (Unexposed, F-)	c	d	c + d
Sum	a + c	b + d	a + b + c + d = N

"F+" — the risk factor is present; "F–" — the risk factor is absent.

AR (_____) = the number of events (good or bad) in treated or control groups, divided by the number of people in that group.

The AR of events in the control group (the group refused treatment — the risk factor is present) = $\mathbf{AR_{(F+)}} = a / (a + b)$.

The AR of events in the treatment group (the risk factor is absent)

= $\mathbf{AR_{(F–)}}$ =

_____ **(absolute risk reduction)** = $\mathbf{AR_{(F+)} – AR_{(F–)}}$

RR (_____) = $\mathbf{AR_{(F+)} / AR_{(F–)}}$

SIGNIFICANT DIFFERENCE

If the RR ____ 1 is a protective factor.
If RR ____ 1, no significant difference between risk and control groups.
If RR ____ 1, significantly more events in a group with a risk factor than in the control group.

_____ **(relative risk reduction)** = $\mathbf{(AR_{(F+)} – AR_{(F–)}) / AR_{(F+)}}$

RRR (%) = (RR – 1) / RR × 100, RRR is shows the proportion of risk factor influence on outcome.

TASKS

PART A

Task 1

Hypothetical cohort study of the one-year recurrence of acute myocardial infarction (MI) among MI survivors with severe systolic hypertension (HTN, ≥ 180 mm Hg) or normal systolic blood pressure (< 120 mm Hg).

Table 1

Findings

Severe Systolic HTN (Groups)	Recurrent MI		Total population
	Present	Absent	
Yes	36	64	100
No	6	94	100
Sum	42	158	200

Task 2

A prospective cohort study was used to assess the association between obesity (defined as BMI ≥ 30 at baseline) and the incidences of cardiovascular disease. Data were collected on participants between the ages of 35 and 65 who were initially free of cardiovascular disease (CVD) and were under observation over ten years. The table below summarizes the findings.

Table 2

Findings

Groups	Incident CVD	No CVD	Total
Obese	92	508	600
Not Obese	30	320	350
Sum	122	828	950

PART B

Calculate and assess the relative risk (RR), attributable risk (AR), relative risk reduction (RRR) of the cohort study.

Task 1

65 out of 100 patients with a family history experienced a myocardial infarction, and 23 out of 50 with no family history also experienced a myocardial infarction.

Task 2

Among 1,560 women who had been smoking before pregnancy, 680 gave birth to children with various developmental disabilities. 389 out of 2,877 non-smoking women gave birth to children with developmental disabilities.

Task 3

65 out of 100 patients with myocardial infarction had a burdened heredity, while there were 23 patients with myocardial infarction out of 50 healthy patients without a burdened heredity.

Task 4

In the group of patients with rheumatism who underwent a course of bicillin therapy (120 people), 13 people fell ill with influenza, and in the group where the course of bicillin therapy was not completed, 50 out of 100 people fell ill.

Task 5

In the treatment of patients with tuberculosis in a group of 263 patients who underwent chemotherapy and pneumothorax, relapses were observed in 15 people, and in the group where only pneumothorax was used, relapses were observed in 19 out of 77 people.

Task 6

Of 79 children who watched TV daily for up to 60 minutes, 37 showed changes in the electrical excitability of the eye, and of 60 children who watched more than 60 minutes a day, 52 showed changes.

Task 7

Of the 1,560 women who smoked during pregnancy, 580 had low birth weight babies, while 620 out of the 2,877 non-smoking women had low birth weight babies.

THE WILCOXON TEST

The Wilcoxon test, which can refer to either the rank-sum test or a version of the signed-rank test, is a _____ statistical test that compares _____ groups. The tests essentially calculate the differences between sets of pairs and analyze these differences to determine whether they are statistically significantly different from each other.

Key takeaways. The Wilcoxon test compares _____ groups.

The purpose of the test is to determine whether two or more sets of pairs are statistically significantly different from each other.

Understanding the Wilcoxon test. The rank-sum test was proposed by an American statistician _____ in a groundbreaking research paper published in 1945. Nonparametric distributions do not have parameters and cannot be defined by an equation like parametric distributions.

The types of questions that the Wilcoxon test can help answer include things like:

Does a particular drug have a health effect when tested on the same people?

This model assumes that the data comes from two matched or dependent populations tracking the same person or group over time or place. The data is also assumed to be continuous rather than _____. Because it is a _____ test, it does not require a specific probability distribution for the dependent variable in the analysis.

MANN–WHITNEY U TEST ASSUMPTIONS

Some key assumptions for Mann–Whitney U Test are detailed below.

The variable being compared between the two groups must be _____ (able to take any number in a range — for example age, weight, height or heart rate). This is because the test is based on ranking the observations in each group.

The data are assumed to take _____ skewed, distribution. If your data are normally distributed, the unpaired Student's t-test should be used to compare the two groups instead.

While the data in both groups are not assumed to be Normal, the data are assumed to be _____ **in shape** across the two groups.

The data should be two randomly selected _____ samples, meaning the groups have no relationship to each other. If samples are paired (for example, two measurements from the same group of participants), then a paired samples t-test should be used instead.

Sufficient **sample size** is needed for a valid test, usually more than ... observations in each group.

Task 1

The duration of treatment of 10 patients diagnosed with pneumonia in the therapeutic department № 1 and 12 patients with the same disease in the department № 2 is presented in the table.

№ department	Duration of treatment, days											
1	7	10	8	5	7	11	6	14	10	12		
2	8	9	11	9	7	14	14	11	12	11	12	14

Determine if there is a statistically significant difference in the duration of pneumonia treatment in those departments.

Task 2

The level of hemoglobin was determined in patients of the main (1) and control (2) groups. The results are presented in the table.

Group	Hemoglobin in the blood, g/l						
1	125	127	140	126	130	124	128
2	141	140	145	150	148	154	147

Determine if there is a statistically significant difference in hemoglobin levels in the main and control groups.

Task 3

In a randomized controlled trial for a new HIV antiretroviral therapy, a pilot study randomly assigned 14 participants to treated or untreated groups. We need to compare the viral load (virus quantity per milliliter of blood) between these groups.

The data are shown below.

Treated (copies/ml)	540	670	1000	960	1200	4650	4200
Untreated (copies/ml)	5000	4200	1300	900	7400	4500	7500

Task 4

Blood cholesterol levels were measured in two groups of patients: the first group served as the control, while the second group followed a diet with reduced animal fat for one month.

The results are presented in the table below.

Group	**Blood cholesterol level, mmol/l**									
1st	6.8	6.3	7.1	7.6	6.7	7.2	6.8	6.9	7.0	7.0
2nd	6.1	6.4	6.5	7.8	5.3	7.0	5.0	5.1	6.0	5.0

FISHER'S EXACT TEST

In statistics, two groups of criteria are distinguished: _____ and nonparametric. Nonparametric statistics are based on the ranking of signs or sign criteria. Such methods include the exact Fisher test.

When the sample size is _____ we can evaluate all possible combinations of the data and compute what are known as exact P-values.

When one of the _____ values (note: not the observed values) in a 2 × 2 table is less than 5, and especially when it is less than 1, in this case Fisher's Exact test, proposed in the mid-1930s almost simultaneously by Fisher, Irwin and Yates, can be applied.

The _____ hypothesis for the test is that there is no association between the rows and columns of the 2 × 2 table, so that the probability of a subject being in a particular row is not influenced by being in a particular column. If the columns represent the study group and the rows

represent the outcome, then the null hypothesis could be interpreted as the probability of having a particular outcome not being influenced by the study group, and the test evaluates whether the two study groups differ in the proportions with each outcome.

An important assumption for all of the methods outlined, including Fisher's Exact test, is that the binary data are _____. If the proportions are correlated then more advanced techniques should be applied. For instance in the leg ulcer if there were more than one leg ulcer per patient, we could not treat the outcomes as independent.

The test is based upon calculating directly the _____ of obtaining the results that we have shown (or results more extreme) if the null hypothesis is actually true, using all possible 2 × 2 tables that could have been observed, for the same row and column totals as the observed data. These row and column totals are also known as marginal totals. What we are trying to establish is how extreme our particular table (combination of cell frequencies) is in relation to all the possible ones that could have occurred given the marginal totals.

$$\text{.....} = \frac{\mathbf{(a+b)! * (c+d)! * (a+c)! * (b+d)!}}{\mathbf{N!}}$$

.....

Task 1

Do the following data allow the use of aspirin to be effective?

Result	Did not take aspirin	Received aspirin treatment
Thrombosis is present	8	1
No thrombosis	1	7

Task 2

Does the following data mean that the use of halothane anesthesia is safer than morphine?

Result	Halothane	Morphine
Alive	7	1
Dead	2	10

Task 3

An outbreak of gastroenteritis was reported in the city of N. Researchers linked the infection to tap water and investigated the correlation between water consumption and the number of cases of infection. What conclusions can be drawn from the data below?

Results	1 cup	4 cups
Number of cases	2	10
The number of non-infected people	12	5

CONTENT

Учебное издание

Мороз Ирина Николаевна
Власова Светлана Викторовна
Куницкая Светлана Васильевна и др.

БИОМЕДИЦИНСКАЯ СТАТИСТИКА

BIOMEDICAL STATISTICS

Практикум

На английском языке

Ответственная за выпуск И. Н. Мороз
Переводчик А. В. Крылова-Олефиренко
Компьютерная вёрстка Н. М. Федорцовой

Подписано в печать 10.07.25. Формат 60×84/16. Бумага писчая «Снегурочка».
Ризография. Гарнитура «Times».
Усл. печ. л. 3,02. Уч.-изд. л. 1,81. Тираж 123 экз. Заказ 489.

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный медицинский университет».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/187 от 24.11.2023.
Ул. Ленинградская, 6, 220006, Минск.